DOI: http://dx.doi.org/10.18203/issn.0019-5359.IndianJMedSci20163531

NOVEL BIOINFORMATICS APPROACH DETECTS HUNDREDS OF PREVIOUSLY UNDETECTED SPLICED TRANSCRIPTS DISCOVERED FROM *CAENORHABDITIS ELEGANS* GENOME

KASHYAP LUV, SHRIVASTAVA KRITI, SHRIVASTAVA AAKRITI[1], UGAM KUMARI CHAUHAN

ABSTRACT

CONTEXT: *With the completion of genome sequence of several organisms including free-living soil nematode Caenorhabditis elegans, precise genome annotations of this sea of raw information are now of prime importance, as they allow the accurate definition of generic regions. Alternative splicing is seen in nearly all metazoan organisms as a means for producing functionally diverse polypeptides from a single gene.* ***AIM:*** *In this study, we performed a detailed and in-depth analysis of the full genomic sequence of one of the six chromosomes of C. elegans.* ***MATERIALS AND METHODS:*** *In this study, several bioinformatics tools including gene/exon prediction programs, ORF finders, blast analysis tools, and alignment programs were used to analyze the genes/exons encoded by chromosome 1 of C. elegans with special reference to alternatively spliced transcripts.* ***CONCLUSION:*** *Using these tools, we have predicted >200 new alternatively spliced hypothetical transcripts from the genes encoded by chromosome 1 in C. elegans. These new spliced transcripts were identified from unusually large untranslated (UTR) regions and large introns present at the 3′ and 5′ ends of the genes with a maximum number of transcripts predicted from 5′ UTR analysis. Further studies and subsequent confirmation of these alternatively spliced transcripts will enhance our understanding of the genome structure, expression, and in elucidating their role during the development of C. elegans.*

Key words: *Alternative splicing, Caenorhabditis elegans, gene/exon prediction, genome analysis*

INTRODUCTION

One of the most remarkable observations stemming from the sequencing of genomes of diverse species is that the number of protein-coding genes in an organism does not correlate with its overall cellular complexity. Alternative splicing, a key mechanism for generating protein complexity, has been suggested as one of the major explanations for this discrepancy between the number of genes and genome complexity. Through alternative splicing, the information stored in the genes of complex organisms can be edited in several ways, making it possible for a single gene to specify two or more distinct proteins. Recent analyses of sequence and microarray data have suggested that alternative splicing plays a major role in the generation of proteomic and functional diversity in almost all metazoan organisms.[1] The nematode *Caenorhabditis elegans* with its rapid life cycle and short lifespan has become a major system for biological study. It is an important, well-studied organism used in biomedical research as a model for human development, genetics, ageing, and diseases. The number of genes in the worm genome is comparable to that in other larger animals, suggesting that although *C. elegans* is small, it has comparable biological complexity to that of other animals. Over half of *C. elegans* genes have human orthologs, while approximately 42% of human disease genes have a homolog in *C. elegans,* allowing the scientist to infer results obtained on them to higher complex organisms including humans.[2,3] The interpretation of the *C. elegans* genome represents the next grand challenge at the interface of computing and biology. It is now well established that alternative splicing is one of the most significant components of the functional complexity of higher eukaryotic genomes such as *C. elegans*, drosophila, mouse, and humans.

Determining the extent and importance of alternative splicing required the confluence of critical advances in data acquisition, improved understanding of biological processes, and the development of fast and accurate computational analysis tools. Several different strategies have been applied to alternative splicing analysis including the following section.

1. EST mapping against mRNA
2. mRNA/EST/protein mapping to the genome
3. Splicing microarray analysis
4. *Ab initio* machine learning approaches.

However, none of the above approaches have been fully successful in delineating all possible alternative splice transcripts of a gene because of their inherent limitations.[4-6] A more complete understanding of alternative splicing requires an unbiased consideration of all alternative mRNA

Department of Environmental Science, Center for Biotechnology Studies, [1]Department of Bioinformatics, Awadhesh Pratap Singh University, Rewa, Madhya Pradesh, India

Address for correspondence:
Dr. Aakriti Shrivastava, Guest Faculty, Department of Bioinformatics, Awadhesh Pratap Singh University, Rewa - 486 001, Madhya Pradesh, India.
E-mail: aakriti_s@yahoo.in

isoforms. Since most of the work and studies have been limited to humans and mouse, not much emphasis has been given to study the alternatively spliced transcripts from *C. elegans* genome. This was the motivational factor to take up this study.

The goal here was to use a novel bioinformatics method capable of delineating all possible spliced transcripts of a gene. Our work comprised complete analysis of the un-annotated unusually large intronic, 5', and 3' untranslated (UTR) genomic regions of chromosome 1 of *C. elegans.* Our major thrust was on finding new exons and genes encoded by chromosome 1 using a combination of bioinformatics tool with a special emphasis on finding novel alternatively spliced transcripts arising from various genes. Around 180-200 new, alternatively spliced transcripts and exons were identified during chromosome one analysis. These new coding sequences in the alternatively spliced transcripts were identified mostly from unusually large UTR regions and large introns present at the 3' and 5' ends of the genes. Furthermore, to experimentally validate our findings, we performed real time-PCR using gene specific primers and RNA isolated from mixed population of *C. elegans* for few of the predicted spliced transcripts of the genes. These new coding sequences, not annotated or identified earlier, will not only add to the available splice database of *C. elegans* but also enhance our knowledge about understanding of the genome structure and evolution of higher eukaryotes especially in context to humans.

MATERIALS AND METHODS

Tools used for computational and bioinformatics analysis

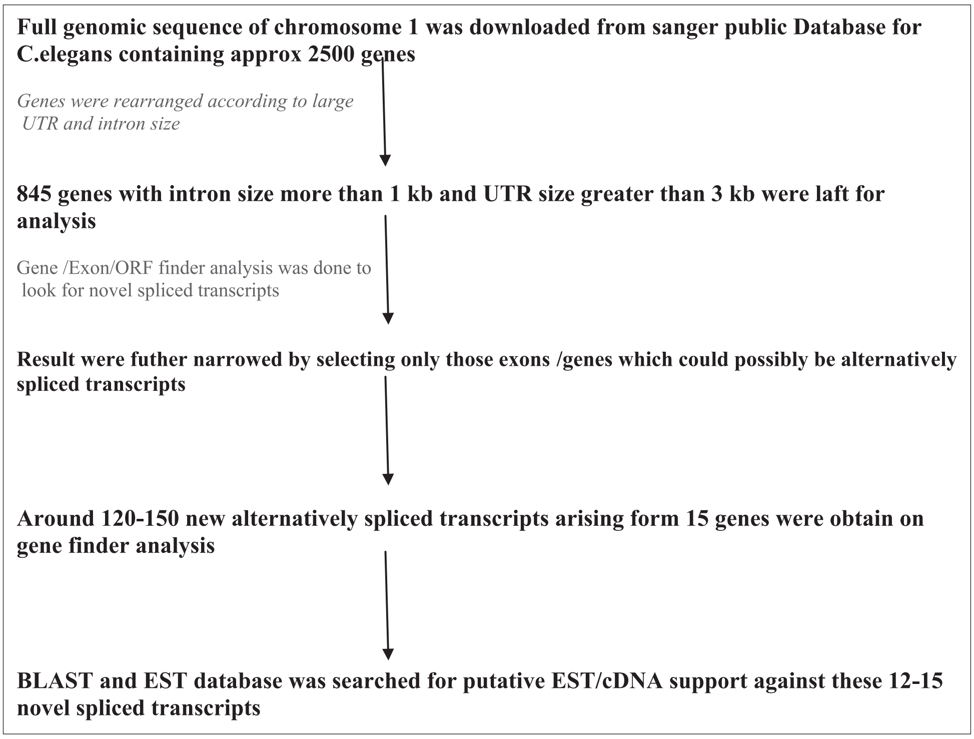
a. Sequence data set: The full genomic sequence data of chromosome 1 of *C. elegans* were downloaded from the *C. elegans* public database located at http://www.sanger.ac.uk/Projects/C_elegas/Genomic_Sequence_htmlhttp://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide: http://www.wormbase.org/db/seq/gbrowse
b. The Gene scan program[7] predicts complete, partial, and multiple genes on both DNA strands. It can be used to identify introns, exons, promoter sites, and polyA signals, among others. It is available at http://genes.mit.edu/GENSCAN.html.
c. The FEX (Find Exon)[8] program initially predicts internal exons by linear discriminant function, evaluating open reading frames (ORFs) flanked by GT and AG base pairs (the 5' and 3' ends of typical introns). It is available at http://www.softberry.com/berry.phtml.
d. The GeneBuilder system[9] is based on a prediction of functional signals and coding regions by different approaches in combination with similarity searches in proteins and EST databases. It is available at http://l25.itba.mi.cnr.it/~webgene/genebuilder.html.
e. The Twinscan is a system[10] for predicting gene structure in eukaryotic genomic sequences. It combines the information from predicted coding regions and splice sites with conservation measurements between the target sequence and sequences from a closely related genome. It is available at http://genes.cse.wustl.edu/.
f. FGENESH[11] is also based on the Hidden Markov Model (HMM). It is available at http://sun1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind.
g. HMM Genefinder[12] is a program for prediction of genes in vertebrate and *C. elegans* genomic sequences. The program is based on an HMM, which is a probabilistic model of the gene structure and is trained using a criterion called conditional maximum likelihood, which maximizes the probability of correct prediction. It is available at http://www.cbs.dtu.dk/services/HMMgene/hmmgene1_1.php.
h. The GeneSplicer is a fast, flexible system for detecting splice sites in the genomic DNA of various eukaryotes.[13] It is available at http://cbcb.umd.edu/software/GeneSplicer/.
i. The program UTR scan[14] looks for UTR functional elements by searching through user submitted sequence data for the patterns defined in the UTRsite collection. It is located at http://www.ba.itb.cnr.it/BIG/UTRScan/.
j. The ORF Finder is a graphical analysis tool that finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. It is available at http://www.ncbi.nlm.nih.gov/projects/gorf/.
k. The splice predictor implements Bayesian models for splice site prediction.[15] It is available at http://bioinformatics.iastate.edu/cgi-bin/sp.cgi.
l. The NetGene2 server[16] is a service producing neural network predictions of splice sites in human, *C. elegans,* and *A. thaliana* DNA. The method is based on a hybrid of the Markov model and neural networks and is available at http://www.cbs.dtu.dk:80/services/NetGene2.

Other bioinformatics tools used are as follows

i. DNA tools: DNA tools were used to unnumber the genomic sequences before entering into the prediction program. It is available at http://biology.semo.edu/cgi-bin/dnatools.pl.
ii. BLAT: BLAT or The BLAST-Like Alignment Tool is designed to quickly find sequences of ≥95% similarity of length ≥40 bases. It is available at http://www.wormbase.org/db/searches/blat.
iii. Yuji Kohara's *C. elegans* EST database at the http://www.ddbj.nig.ac.jp/c-elegans/html/CE_INDEX.html.

Approach used for detecting novel spliced transcripts from *C. elegans* genome

Our approach for in-depth analyses and for identifying novel



Full genomic sequence of chromosome 1 was downloaded from sanger public Database for C.elegans containing approx 2500 genes

Genes were rearranged according to large UTR and intron size

845 genes with intron size more than 1 kb and UTR size greater than 3 kb were laft for analysis

Gene /Exon/ORF finder analysis was done to look for novel spliced transcripts

Result were futher narrowed by selecting only those exons /genes which could possibly be alternatively spliced transcripts

Around 120-150 new alternatively spliced transcripts arising form 15 genes were obtain on gene finder analysis

BLAST and EST database was searched for putative EST/cDNA support against these 12-15 novel spliced transcripts

Figure 1: Flowchart depicting our approach

alternatively spliced transcripts of genes consisted of the following major steps [Figure 1]:

1. Downloading the genome sequence data and arranging genes: We started our work by downloading the complete intronic and UTR data information available about chromosome 1 of *C. elegans* from the Sanger Center Public Database and other relevant databases (as mentioned in Materials and Methods).
2. Rearrangement and selection of potential genes for computational analysis: These data were further screened and scaled down in accordance to the size of large introns and UTR regions. Here, our aim was to identify genes with unusually large size of 5' and 3' UTR and intron regions. Thus, we were left with around 875 genes having potentially large intronic and 5' and 3' UTR regions.
3. Gene/Exon/ORF Finder analysis: The next task was to explore these large gap regions in the 875 potential genes. Our analysis started with running these unusually large gaps at intronic, 5', and 3' UTR regions on a pre-selected order of tools gene/exon/OFR finding and several other bioinformatics analysis tools. These programs combine a variety of gene/exon prediction methodologies including *ab initio* predictions (Gene finder, FGENESH), EST, protein-based comparisons (Ensembl), sequence conservation metrics (TWINSCAN), and many more (as mentioned in Materials and Methods). Therefore, the results obtained were much authenticate then those obtained using a single gene/exon predicting tool. When large genomic sequences from large UTR and intron gaps were fed into these tools, they predicted several exons possibly capable of replacing the existing exon(s) and thus creating alternatively spliced variant of a gene.
4. From the several exons predicted above, we selected only the "common exons" capable of replacing the existing exon(s), and thus, creating spliced transcript of the gene without affecting the reading frame of the protein. Furthermore, the possibility of occurrence of that spliced exon/transcript was analyzed by a comparative analysis between the original protein and the new protein formed by addition/exclusion of alternative exons using various alignment tools. Finally, several other parameters such as percent-amino acid replacement, codon usage, sense nature (i.e., whether from positive or negative strand), and the probability score of occurrence of that exon were also checked to ensure the accuracy of predicted spliced transcript of the gene. Thus, using the above approach, we were successful in detecting splicing in 160–180 genes and giving rise to possibly 200–250 new, alternatively spliced transcripts/exons from chromosome 1 of *C. elegans.*
5. Homology, BLAST, and EST analyses: Following the computational predictions of these novel spliced transcripts, the Yuji Khoara's and NCBI dbEST *C. elegans* EST database and various other relevant databases were searched to look for putative EST/cDNA support for the possible occurrence of these new exons/transcripts. First, using various alignment tools, we looked for insertions or deletions in ESTs relative to a set of known mRNAs or by aligning the ESTs exactly to their identified genomic sequence in the draft genome to identify potential alternative splices. Second, as intronic sequences at splice junctions are highly conserved (99.24% of introns have a GT-AG at their 5' and 3' ends, respectively), so these splices were identified and intronic splice junction donor and acceptor

sites were checked using various splice site predicting tools. Finally, NCBI BLAST search was performed to look for homology or prospective similarity with other polypeptides of these new spliced transcripts.

Experimental validation of computationally predicted spliced transcripts

After the failure in search for supporting EST/cDNA matches for our new prediction, the next way to confirm our findings was to validate them in lab using any suitable method. RT-PCR is one of the most powerful and directs methods to detect transcript variants due to alternative splicing. The RT-PCR is easier and more popular than microarray techniques in terms of confirmation of alternatively spliced variants of an individual gene. We selected 10 genes randomly from the total set of genes for which we had predicted novel spliced transcripts using our new bioinformatics methodology as detailed above. We performed RT-PCR using gene specific primers and RNA isolated from mixed population of *C. elegans* for these selected genes. RT-PCR validation confirmed the existence of the computationally predicted transcripts for five out of the total ten selected genes. Detail results for which are either already published, e.g., *lfe-2* encoding C46H11.4 gene,[5] RhoGEF domain encoding Y95B8A.12 gene,[6] cadherin encoding W02B9.1 gene,[4] and few more that are yet to be published.

RESULTS AND DISCUSSION

Our aim was to use a novel bioinformatics approach involving the use of a wide array of bioinformatics tools and programs capable of delineating all possible spliced transcripts of a gene. We roughly analyzed around 857 genes having unusually large size of 5' and 3' UTR and intron regions and have identified 160-180 genes in which splicing occurs. The total number of transcripts arising from alternative splicing in these genes was found to be around 180-200 new, alternatively spliced transcripts/exons from chromosome 1 of *C. elegans* [Figure 1]. Comprehensive lists of genes in which we found new spliced transcripts are given in Figures 2 and 3.

These new coding sequences in the alternatively spliced transcripts were identified from unusually large UTR regions and large introns present at the 3' and 5' ends of the genes. The maximum number of unreported new exons were predicted from 5' UTR, followed by 3' UTRs and introns [Figure 2]. All new exons were capable of splicing with the existing gene products generating new splice variants as the splice junctions were conserved at the splice donor and acceptor sites. A number of new, unreported possible splice variants predicted accounted a total of 196 including maximum from 5' UTR (n = 108), followed by 3' UTR (n = 63), introns present toward 5' end (n = 17), and 3' end (n = 8) [Figure 3].

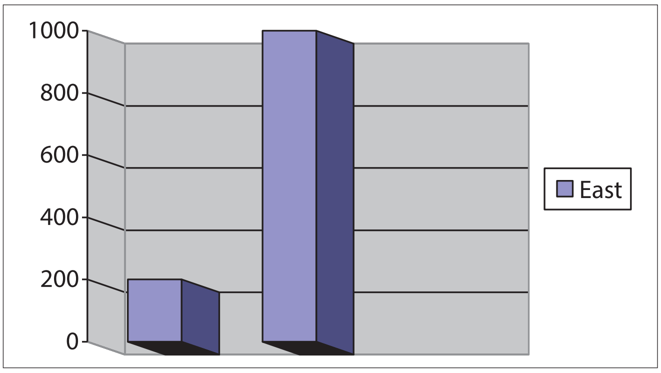


Figure 2: A comparative study between the number of genes analyzed and number of genes having predicted splice transcripts: A comparison between total number of genes analyzed (having potentially large unusually large untranslated and intronic gaps) in our study and the number of genes having predicted spliced transcripts based on bioinformatics analysis

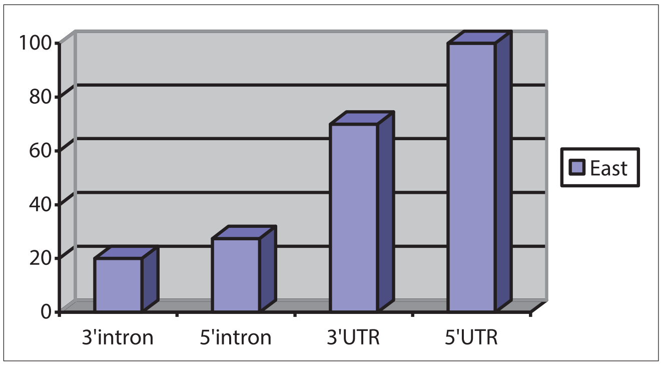


Figure 3: Predictions of new exons from the selected regions: A comparative picture between gap regions having predicted spliced transcripts. The maximum number of unreported new exons were predicted from 5' unusually large untranslated (n = 108), followed by 3' unusually large untranslated (n = 63), and intronic gap regions 5' end (n = 17) and 3' end (n = 8) from the total number of genes having predicted spliced transcripts (n = 196)

Our results demonstrate that we are still far from completely deciphering these hidden transcripts from the genome of sequenced organism, and most of the studies have probably underestimated the extent of alternative splicing. Thus, the end goal of alternative splicing annotation is to identify and catalog all mRNA transcripts in the cells and develop an exhaustive catalog of alternative transcripts of an organism to fully understand the complexity of eukaryotes. Although we were successful in identifying potentially new spliced transcripts and alternative exons from chromosome 1 of *C. elegans,* our findings indicate that there could be approximately 1000 or more alternatively spliced transcripts expressed from the genome of *C. elegans* that have not been annotated or identified earlier. This could be the reason why the number of gene products is suspected to be underestimated. With this experience, we propose that the genome data may be analyzed using combination of bioinformatics tools and programs to predict the full repertoire of a gene product. These new coding sequences and transcripts, not annotated or identified earlier, will be helpful to the biological community in several ways: First, they will not only help in increasing the available database for alternatively spliced genes in *C. elegans* but also in pointing toward the complex mechanism of alternative splicing in *C. elegans* genes and their role in downstream regulatory steps. Second, their findings indicate the urgent need to develop the more efficient algorithms and methods capable of identifying the full catalog of alternative spliced transcripts of a gene. Moreover, similar exhaustive studies could be taken up in several other finished genomes, especially of humans with whom *C. elegans* share a close gene homology. Finally, due to limited domain of our work, further studies using more advanced techniques such as the RNA interference (RNAi) could be taken up, which would enhance our knowledge about the biological and functional significance of these spliced transcripts and their possible role in *C. elegans* gene working and regulation.

REFERENCES

1. Mironov AA, Fickett JW, Gelfand MS. Frequent alternative splicing of human genes. Genome Res 1999;9:1288-93.
2. Ahringer J. Turn to the worm! Curr Opin Genet Dev 1997;7:410-5.
3. Culetto E, Sattelle DB. A role for *Caenorhabditis elegans* in understanding the function and interactions of human disease genes. Hum Mol Genet 2000;9:869-77.
4. Kashyap L, Tabish M. Alternatively spliced isoforms encoded by cadherin genes from *C. Elegansgenome*. Bioinformation 2007;2:50-6.
5. Kashyap L, Tabish M, Ganesh G, Dubey D. Computational and molecular characterization of multiple isoforms of lfe-2 gene in nematode *C. Elegans*. Bioinformation 2007;2:17-21.
6. Kashyap L, Tabish M, Ganesh S, Dubey D. Identification and comparative analysis of novel alternatively spliced transcripts of RhoGEF domain encoding gene in *C. Elegans* and *C. Briggsae*. Bioinformation 2007;2:43-9.
7. Burge C. Identification of Complete Gene Structure in Human Genomic DNA. PhD Thesis. Stanford, CA: Stanford University; 1997.
8. Solovyev VV, Salamov AA, Lawrence CB. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucleic Acids Res 1994;22:5156-63.
9. Milanesi L, D'Angelo D, Rogozin IB. GeneBuilder: Interactive in silico prediction of gene structure. Bioinformatics 1999;15:612-21.
10. Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouzé P, Brunak S. Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. Nucleic Acids Res 1996;24:3439-52.
11. Resch A, Xing Y, Modrek B, Gorlick M, Riley R, Lee C. Assessing the impact of alternative splicing on domain interactions in the human proteome. J Proteome Res 2004;3:76-83.
12. Krogh A. Two methods for improving performance of an HMM and their application for gene finding. In: Gaasterland T, editor. Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology. Menlo Park, CA: AAAI Press; 1997. p. 179-86.
13. Pertea M, Lin X, Salzberg SL. GeneSplicer: A new computational method for splice site prediction. Nucleic Acids Res 2001;29:1185-90.
14. Modrek B, Lee C. A genomic view of alternative splicing. Nat Genet 2002;30:13-9.
15. Brendel V, Xing L, Zhu W. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinformatics 2004;20:1157-69.

How to cite this article: Luv K, Kriti S, Aakriti S, Chauhan UK. Novel bioinformatics approach detects hundreds of previously undetected spliced transcripts discovered from *Caenorhabditis elegans* genome. Indian J Med Sci 2016;68:49-53.

Source of Support: Nil. **Conflict of Interest:** None declared.